DOCUMENT RESUME

ED 360 335 TM 020 132

AUTHOR Webb, Lynn C.; And Others

TITLE The Use of Cognitive Taxonomies in Licensure and Certification Test Development: Reasonable or Customary?

PUB DATE Apr 93

NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Allied Health Occupations; *Allied Health Personnel; *Classification; *Cognitive Processes; Factor Analysis; Higher Education; *Licensing Examinations (Professions); Multiple Choice Tests; Scores; *Test Construction; Test Items; Thinking Skills

IDENTIFIERS *Blooms Taxonomy; Hierarchical Models; *Test Specifications

ABSTRACT

The functioning of a cognitive taxonomy within the test specifications of an allied health certification examination was studied. The taxonomy used was a simplification of the scheme of B. S. Bloom (1956), in which items were classified as comprehension, application, or analysis. Whether items written purposely to assess higher order cognitive processes actually assessed differing levels of cognitive processing was explored. A factor analysis of responses of 627 examinees does not support a cumulative hierarchical model of cognitive complexity. Several cases of model misfit were observed, in which some examinees performed better on the higher level subtest than on the lower level subtest, a finding that is counter to that which would be predicted under a functioning cumulative, hierarchical model. A finding that supported the hypothesis of functioning cognitive levels was that examinees who scored in the upper quartile of the higher level subtest were more likely to pass the examination than were those who scored in the lowest quartile. Overall, results support continued use of a cognitive classification dimension for test specifications. Implications for test specifications development, test construction, item writing, and score reporting are presented, as are limitations and suggestions for further research. Five tables present study findings. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED360835

The Use of Cognitive Taxonomies in Licensure and Certification Test Development: Reasonable or Customary?

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
LYNN C. WEBB
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the
American Educational Research Association
Atlanta, GA
April 1993

by

Lynn C. Webb, EdD
Assistant Director
Health Programs Dept.
American College Testing

Gregory J. Cizek, PhD
Assistant Professor of
Educational Research
University of Toledo

John C. Kalohn, PhD
Program Associate
Health Programs Dept.
American College Testing

Correspondence regarding this research should be directed to Lynn C. Webb, EdD, American College Testing, P.O. Box 168, Iowa City, IA, 52243.

BEST COPY AVAILABLE

ABSTRACT

This research addressed the functioning of a cognitive taxonomy within the test specifications of an allied health certification examination. The cognitive taxonomy studied was a simplification of the Bloom's (1956) general scheme, in which items were classified as Comprehension, Application, or Analysis. The research investigated whether test items written purposefully to assess the higher order cognitive processes actually assessed differing levels of cognitive processing.

Factor analysis of examinees responses did not provide support for a cumulative hierarchical model of cognitive complexity; instead, only one factor emerged. Several cases of model misfit were also observed, in which some examinees performed better on the higher level subtest (Analysis) than on the lower level subtest (Comprehension)--a finding that is also counter to that which would be predicted under a functioning cumulative, hierarchical model.

A finding that supported the hypothesis of functioning cognitive levels was that examinees who scored in the upper quartile of the higher level subtest (i.e., Analysis) were more likely to pass the examination than those examinees who scored in the lower quartile of that subtest.

Overall, the results yielded qualified support for the continuing use of a cognitive classification dimension for test specifications. Implications of the research for test specifications development, test construction, item writing, and score reporting are presented. Limitations and suggestions for future research are also provided.

The Use of Cognitive Taxonomies in Licensure and Certification Test Development: Reasonable or Customary?

Cognitive taxonomies are widely used as one dimension in delineating test specifications for licensure and certification testing programs. Some common reasons for incorporating cognitive taxonomies into test specifications are to ensure that "higher order" cognitive processes are assessed, or to promote a match between test items and complex job/task demands.

The most common cognitive classification system in current use is that presented by Bloom (1956), or some simplification of Bloom's general scheme. The Bloom taxonomy suggests that cognitive functioning can be represented with a hierarchical structure from lowest level of functioning (Knowledge or Recall) to higher, more complex, or more sophisticated levels, such as Application, Analysis, Synthesis, or Evaluation. The cumulative hierarchical structure of cognitive functioning presented in the taxonomy rests on the assumption that "simpler behaviors may be viewed as components of the more complex behaviors" (Bloom, 1956, p. 16).

Background

Sparse empirical work has been initiated to validate the Bloom taxonomy or its variations, to verify the existence of the asserted levels, or to support its application for the uses noted above. We know of no research on this topic conducted in the area of licensure and certification testing. Accordingly, we concur with Madaus, Woods, & Nuttall (1973) who observed that:

> "Given the widespread use of the [Bloom] Taxonomy in formulating objectives in a multitude of curricular areas, for various types of students at differing levels of education, further investigation of the Taxonomy's assumptions would not be without considerable practical value" (p. 262).

What work has been done has yielded mixed results. For example, an investigation by Kropp and Stoker (1966) studied high school students' performance on science and social studies tests and provided support for a cumulative, hierarchical taxonomic structure involving the first four of Bloom's levels (i.e,. knowledge, comprehension, application, analysis). In their research, however, the synthesis and evaluation level items did not perform as would be predicted. Further, they also observed a pattern of increasing correlations between subtest scores (i.e., taxonomic levels) and scores on a test of reasoning ability as taxonomic level increased. Thus, although support for the presumed taxonomic structure was obtained, some influence of a "general mental ability" construct was also observed.

In a subsequent reanalysis of Kropp and Stoker's data, Madaus, Woods, and Nuttall used a causal modeling approach to ascertain the existence of direct and indirect links between levels of the taxonomy. They found "a decline in the magnitude of the direct links between adjacent [taxonomic] levels as the levels became extremely complex and...numerous indirect links between nonadjacent levels" (p. 261). Additionally, they noted that only one indirect link (between the comprehension and analysis levels) remained when a "g" factor of general mental ability was introduced in to the causal model (p. 261).

Finally, a study was conducted by Little (1971) involving preservice elementary education students' performance on an examination comprised of subtests designed to assess each of the six taxonomic levels. Analysis of correlations between the Knowledge, Comprehension, Application, and Analysis subtests of the examination supported the existence of a hierarchy for these four levels, but failed to support the stated hierarchy composed of a six-level hierarchical clustering scheme.

Applications in Certification and Licensure Testing

Cumulative hierarchical models of cognitive functioning--most commonly,

simplifications of the Bloom (1956) model--appear to be widely relied upon in licensure and certification testing programs. This reliance is observable in the frequent use of cognitive taxonomies in test specifications development, test item writing, and test score (or subscore) reporting to examinees. For example, in role delineation studies, cognitive levels are sometimes incorporated into the survey instrument, as one of three dimensions of interest. The three dimensions can be represented as: 1) FREQUENCY - the frequency with which a task or skill is necessary in practice; 2) CRITICALITY - the judged relationship between proper performance of the task or skill and safe or effective practice; and 3) COGNITIVE - the level of cognitive processing required by the practitioner to properly perform the task. This third dimension has also been called a "Complexity" dimension and is described by Cavanaugh (1991):

> "The Complexity scale is designed to estimate the level of cognition required to perform each task. This information provides a basis for matching the level of complexity for assessment with the level of complexity required in performance on the job" (pp. 31-32).

We agree with Cavanaugh that the most appropriate point in the test development process for incorporating a cognitive dimension is at the beginning (i.e., during task analysis or role delineation). However, we observe that task analyses often focus on the aspects of frequency and criticality with consideration of cognitive levels reserved for the item development phase, in which fairly arbitrary percentages are assigned to cognitive dimensions represented in test specifications. Although there may be a strong logical rationale for this approach, it provides little empirical support for the model of practice hypothesized or for the validity of test score interpretations.

In summary, our review of the literature indicates that, in general, the cumulative, hierarchical cognitive functioning model lacks strong empirical support. Further, although a cumulative hierarchical structure of cognitive functioning is often used in licensure and certification testing programs, the existence of these functioning cognitive levels is often only presumed. That is, the implicit assumption of entities responsible for developing and administering the programs is that successful performance on test items designed to assess higher level cognitive functioning are better indicators of content mastery. However, the tenability of this assumption has not been fully explored.

Thus, this research investigates whether test items written purposefully to assess the higher order cognitive processes actually assess differing levels of cognitive processing. Specifically, it is hypothesized that higher levels of performance on "higher order" item groupings should be associated with greater success on the total test (measured either in terms of total test score or pass/fail classifications). Further, it is hypothesized that, if the hierarchical structure of cognitive levels exists, performance on subtests defined according to cognitive levels should reflect that hierarchy.

Procedures

Data for this research were collected as part of the annual administration of a 200-item certification examination for candidates in an allied health field[1]. The examination blueprint specified test construction procedures utilizing the common three-dimensional matrix. One dimension of the matrix describes content categories; a second dimension specifies cognitive classification; the third level indicates frequency (i.e., the number of test items per cell). The cognitive classification system employed

[1] On this 200-item test form, one item was double-keyed and one item was scored correct for all examinees. Thus, a total of 198 items were used for this analysis.

was a simplification of the Bloom taxonomy, using three levels of classification (COMPREHENSION, APPLICATION, and ANALYSIS). The numbers of items allocated to each of these categories were 45, 117, and 36, respectively. Responses were obtained from 627 examinees to traditional five-option multiple-choice items in a 1992 administration of the examination.

Two strategies were used to identify possible existence of functional cognitive process classifications in the test items. First, factor analytic methods were employed to discern the number of factor(s) assessed by the test. The hypothesis of interest was: If cognitive complexity of the test items is a differentiating factor, distinct factors identifying the levels should emerge.

Second, data analysis consisted of obtaining overall proficiency estimates for each examinee. Initially, an Item Response Theory (IRT) approach was attempted to obtain the overall proficiency estimates. However, the IRT approach proved unworkable; consequently, total test scores were utilized as substitute measures of overall ability level[2]. Subtest scores (defined by cognitive classifications) and pass/fail decisions were also recorded for each examinee. Total test scores were correlated with subtest scores to reveal the extent to which "higher-level subtest" scores are associated with higher examinee abilities. Also, the frequency of examinees with "aberrant" response patterns (i.e., low scores on lower-level subtests and high scores on higher-level subtests) who pass the examination was examined using contingency table analysis.

Results

Correlations between subtest scores and total test scores, as well as numbers of items in each subtest and total test are presented in Table 1. As would be expected, the correlations were all high, positive, and significantly

[2] Because the relationship between IRT ability estimates using the Rasch Model and total raw scores is frequently observed to yield correlations near +1.0, the use of total raw scores here seems reasonable.

different from zero at $p<.001$. The largest correlation is seen for the subtest with the greatest number of items in common with the total test. Table 2 provides sets of subtest intercorrelations. The lower triangle gives the unadjusted correlations between subtests, while the upper triangle contains disattenuated subtest intercorrelations (subtest reliabilities appear in parentheses on the diagonal). Table 2 reveals that subtest scores are highly correlated (again, all intercorrelations were significant at $p<.001$); further, the corrected correlations all approach +1.00 (i.e., true scores on the subtests are nearly perfectly correlated), providing some support for the hypothesis that the subtests may be measuring a unitary construct.

Exploratory factor analysis results were also consistent with this hypothesis. Examination of the Pearson interitem correlations revealed a fairly uniform matrix of small correlations. Thus, it was decided to utilize an alternative similarity coefficient for dichotomous variables as input for the factor analysis, and the Jaccard index (see Kotz, 1985, p. 399) was selected in order to increase observed variability. An initial analysis was conducted without limiting the number of factors to be extracted. Final analysis, however, constrained the number of factors estimated to five.

The unrotated factor analysis solution revealed the variance explained and percentages of total variance explained by the factors shown in Table 3. Application of Kaiser's criterion (Kaiser, 1974) suggested retaining three factors. However, the variance explained by the five factors extracted and their corresponding percentage of total variance explained strongly supported the hypothesis of a single primary factor. In an attempt to further simplify the factor structure, a varimax rotation was employed. These results are also presented in Table 3 and are consistent with a single factor interpretation for the structure of the test.

Finally, a contingency table analysis was conducted to determine if differential performance on cognitive subtests was related to pass/fail status. Maximum differentiation was achieved by comparing performance on the

two subtests hypothesized to be most cognitively different under the cumulative hierarchical model (i.e., the Comprehension and Analysis subtests). Two nominal variables, HIGHERCOMP and HIGHERANAL, were created for use in this analysis. Examinees whose percent correct score on the Comprehension subtest was higher than their percent correct score on the Analysis subtest were assigned a value of "1" on the variable HIGHERCOMP; those whose Comprehension percent correct score was lower received a "0". Examinees whose percent correct score on the Analysis subtest was higher than their percent correct score on the Comprehension subtest were assigned a value of "1" on the variable HIGHERANAL. HIGHERCOMP and HIGHERANAL represented the two levels of a cognitive complexity variable in a 2 x 2 contingency table; examinees' PASS or FAIL status on the total test was used for the two levels of the second variable.

Raw data for the 627 examinees and the chi-square test for independence between subtest cognitive complexity and pass/fail status are presented in Table 4. Of the 627 examinees, 500 (79.7%) passed and 127 (20.3%) failed. Regarding examinee performance of the cognitive subtests, a majority of examinees (69.7%) scored higher on the Comprehension subtest than they did on the Analysis subtest; conversely, 190 (30.3%) scored higher on the Analysis subtest than they did on the Comprehension subtest.

A chi-square test resulted in rejection of the null hypothesis of independence between subtest performance and pass/fail status (χ^2 = 19.14, $p<.001$). Although examinees generally performed better on the Comprehension items than on the Application items, those examinees who performed better on the higher level subtest (i.e., Application) compared to the lower level subtest (i.e., Comprehension) were significantly more likely to pass the examination (91% compared to 75%).

A final analysis to investigate the relationship between pass/fail status and cognitive level utilized a contingency table approach. The distribution of examinees' scores on the Analysis subtest was divided into

Lower, Intermediate, and Upper quartiles and is presented in Table 5. A chi-square test resulted in the rejection of the null hypothesis of independence between performance on the Analysis subtest and pass/fail status (χ^2 =296.67, p<.0001). This result indicates that, in general, examinees who performed in the upper quartile of the Analysis subtest were more likely to pass the examination than those who scored in the lower quartile on that subtest.

Discussion

Our results yield fairly consistent, though tentative, interpretations. Analysis of correlations showed that subtests intended to assess differing levels of cognitive processing were highly related. Factor analytic procedures also suggested that variability in performance could be attributed to a single factor; distinct cognitive level factors corresponding to subtest identities did not emerge.

Thus, the results of our preliminary analyses indicate that the cognitive classification system used in the testing program studied does not function as would be expected if well-differentiated, hierarchical levels existed. In the following sections, we emphasize the caution with which our findings should be interpreted, and provide interpretations and suggestions for the future.

Cautions and Limitations

First, this study concerned a single allied health certification testing program with three levels of cognitive complexity. In order to assess the generalizability of our findings, we intend to replicate this research with other licensure and certification programs, using various categorization systems for incorporating cognitive levels. We urge others to attempt replications of this investigation as well.

Second, we recognize that our findings are not unambiguous. For example, we observed that better performance on the more cognitively complex

subtest (i.e., Analysis) was related to success on the test as a whole (i.e., to passing).

Third, it should be noted that this investigation did not attempt to validate the categorization of items comprising the subtests. That is, we did not verify the judgments of committee of content experts who classified the test items according to cognitive level; nor were we able to review the training procedures provided to item writers in order to assess the faithfulness with which they captured the intended cognitive level.

Recommendations

Despite these qualifications, we believe that the findings of this research are both significant and somewhat controversial. This research has important implications for test development practice. First, our research reconfirms the need to investigate the applicability of cognitive levels for licensure and certification testing programs and emphasizes that, if appropriate, empirically-derived levels are desirable. Accordingly, we again note our concurrence with Cavanaugh's (1991) recommendation that the decision to incorporate cognitive levels be based in task or job analysis data. The decision to include cognitive levels should not be made, essentially, as an arbitrary afterthought during the test specifications development phase.

Because it appears that few job analyses consider cognitive levels *a priori*, we recommend that additional research to validate their use be conducted by entities responsible for licensure and certification testing programs. We envision that a review of research regarding the role of cognitive levels in test development and established guidelines for their use would be a welcome addition to the literature on licensure and certification testing.

Second, although our research failed to find evidence for functioning, cognitive levels for the testing program studied, we do not imply that current, rationally-derived cognitive taxonomies are of little use. To the

contrary, it is noted that the incorporation of even non-functioning rationally-derived levels can yield substantial practical benefits. For example, the use of cognitive levels holds obvious benefits for the item-writing process: experience has shown that item writers who lack training in the generation of "higher order" items tend to produce low quality items assessing the lowest levels of cognitive processing. Undoubtedly, the ubiquitous attention paid to cognitive levels during item-writer training has had a generally beneficial effect on the overall quality of licensure and certification test.

Second, entities responsible for credentialling decisions accrue the incidental benefit of increased validity accompanying the use cognitive levels when that use results in expanding and ensuring breadth and depth of content coverage.

Finally, examinees probably benefit from the incorporation of cognitive levels in the licensure and certification processes. The representation of important content in examinee handbooks, candidate guides, etc., can serve as an aid to examinees in test preparation, in developing conceptual schema to represent important content, and in becoming familiar with a framework for organizing relevant information about professional practice that is shared by experts in the field.

REFERENCES

Bloom, B.S. (Ed.) (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain. New York: McKay.

Cavanaugh, S.H. (1991). Response to a legal challenge: Five steps to defensible credentialling examinations. Evaluation and the Health Professions, 14(1), 13-40.

Kaiser, H.F. (1974). An index of factorial simplicity. Psychometrika, 35, 31-36.

Kotz, S. (1985). Encyclopedia of statistical science (Vol. 5). New York: Wiley.

Kropp, R.P. & Stoker, H.W. (1966, February). The construction and validation of tests of the cognitive processes as described in the taxonomy of educational objectives. Research report, Institute of Human Learning, Florida State University (ED 010 044).

Little, R.A. (1971). A Taxonomic Approach to Measuring Achievement in Mathematics 223 - Geometry for Elementary Teachers. Doctoral Dissertation, Kent State University (University Microfilms Order No. 72-15,945).

Madaus, G.F., Woods, E.M., & Nuttall, R.L. (1973). A causal model analysis of Bloom's taxonomy. American Educational Research Journal, 10(4), 253-262.

Waugh, R. (1975). Bloom's taxonomy and mathematics teaching. Australian Mathematics Teacher, 31(6), 209-213.

TABLE 1

Subtest-Total Test Correlations and Numbers of Items

(Based on n=627 Examinees)

Variables	Number of Items	Correlation
Comprehension, Total Test	45 198	.865 (p<.001)
Application, Total Test	117 198	.979 (p<.001)
Analysis, Total Test	36 198	.895 (p<.001)

TABLE 2

Subtest Intercorrelations, Reliabilities, and Adjusted Intercorrelations

(Based on n=627 Examinees)

	SUBTESTS		
	Comprehension	Application	Analysis
Comprehension	(.709)	.977*	.946*
Application	.780	(.899)	.994*
Analysis	.701	.829	(.775)

Notes: 1) Diagonal entries in parentheses are KR-20 subtest reliabilities; uncorrected correlations appear below diagonal; correlations above diagonal (indicated with asterisks) are corrected for attenuation.

2) All correlations significantly different from zero at p<.001.

TABLE 3

Factor Analysis Results

	-----Unrotated Solution----		-----Varimax Rotation----	
Factor	Variance Explained	Percent of Total Variance Explained	Variance Explained	Percent of Total Variance Explained
1	109.69	55.40	71.42	36.07
2	4.00	2.02	38.34	19.36
3	1.07	0.54	1.91	0.97
4	0.85	0.43	3.08	1.56
5	0.62	0.31	1.48	0.75

TABLE 4

Contingency Table Analysis of Subtest Cognitive Level and Pass/Fail Status

		Cognitive Complexity		
		Highercomp	Higheranal	Total
Total Test Pass/Fail Status	FAIL	109 (17.4%)	18 (2.9%)	127 (20.3%)
	PASS	328 (52.3%)	172 (27.4%)	500 (79.7%)
	Totals	437 (69.7%)	190 (30.3%)	627(100.0%)

$\chi^2 = 19.14$, $p<.001$

TABLE 5

Analysis of Performance on Analysis Subtest and Pass/Fail Status

Distribution of Analysis Subtest Scores

	Lower Quartile	Inter Quartile Range	Upper Quartile	Totals
FAIL	101	26	0	127
PASS	42	289	169	500
Totals	143	315	169	627

$\chi^2 = 295.67$, $p<.0001$